White Paper

NEH HK-50181-14

Archive What I See Now: Bringing Institutional Web Archiving Tools to the Individual Researcher

PI: Michele C. Weigle (ODU)

Co-PIs: Michael L. Nelson (ODU), Liza Potts (Michigan State)

Old Dominion University Research Foundation

December 2017

Introduction

Much of our cultural discourse occurs primarily on the Web and its preservation is a fundamental pre-condition for research in history, sociology, political science, media, literature, and other related disciplines. Archiving web pages is an essential method for ensuring these resources are available for future research. This is not unlike the need for techniques to preserve physical versions of historical documents. Institutional archiving tools may help in some cases, but are difficult to install and use. In other cases (social media, interactive webpages), institutional archiving tools may not be able to archive the page as the researcher views it in her web browser.

The goals of this project were two-fold: 1) to enable users to generate files suitable for use by large-scale archives with tools as simple as the "bookmarking" or "save page as" approaches that they already know, and 2) to enable users to access the archived resources in their browser. We focused on bringing the power of institutional web archiving tools like Heritrix (used for crawling and archiving webpages) and *wayback* (used for accessing archived webpages) to humanities scholars through innovative open-source tools for personal-scale web archiving.

Our project proposed building three open-source tools to support this personal-scale web archiving:

- WARCreate – A Google Chrome browser extension that can create an archive of a single webpage in the standard web archiving format and save it to local disk. It can allow a user to archive pages behind authentication or that have been modified after user interaction.
- WAIL - A stand-alone application that provides one-click installation and GUI-based configuration of two popular institutional archiving tools on the user's personal computer.
- Mink - A Google Chrome browser extension that provides access to archived versions of live webpages, including local archives created by WARCreate.

These tools also allow users to upload generated archive files to a specified server for later access. With these tools, collaborating scholars could upload their archives to a common server to create special-purpose collections of various topics.

Our innovation is in allowing individuals to "archive what I see now". The user can create a standard web archive file ("archive") of the content displayed in the browser ("what I see") at a particular time ("now").

Project Activities

This Digital Humanities Implementation Grant was a follow-on to our previous Digital Humanities Start-Up Grant (NEH HD-51670-13).

Our project involved the continuing development of the following tools:

- WARCreate - A browser extension that can create an archive of a single webpage in the standard WARC format and save it to local disk. It can allow a user to archive pages behind authentication or that have been modified after user interaction.
- WAIL (Web Archiving Integration Layer) – A stand-alone application that provides easy installation and GUI-based configuration of both Heritrix and *wayback* on the user's personal computer.
- Mink - A browser extension that provides access to archived versions of live webpages.

The main tasks that we proposed were as follows:

- **Task 1** - New WARCreate Modes. We will extend WARCreate to include the record, countdown, and event modes. We will also re-evaluate and refine the sequential archiving option.
- **Task 2** – WARC File Transfer Methods. With the help of Archive-It, we will create a process to upload user-generated WARCs to a private Archive-It collection. In addition, we will test other secure methods of transporting WARCs between cooperating researchers, such as email and Dropbox (or Google Drive, or other such tool).
- **Task 3** - WAIL User Interface. We will continue to develop the user interface of WAIL, ensuring that the installation process and the use of all options are intuitive.

- **Task 4** - New Wayback Feature Integration into WAIL. We will integrate new features of *wayback* that we develop in other projects into the version included with WAIL.
- **Task 5** – Firefox Version of Mink. We will develop a Firefox version of Mink that will look and operate similar to the Google Chrome version.
- **Task 6** – Mink User Interface. We will further develop the Mink user interface based on feedback from the advisory board. This includes adding the number of archived pages to the icon at the bottom of the page and allowing users to set preferences on how to view a large set of mementos.
- **Task 7** – Local Archive Query in Mink. We will implement the ability for Mink to query a local archive before querying public Memento aggregators.
- **Task 8** – Memento Damage Assessment. Those creating WARCs need to ensure that what they intended to be archived was actually archived. We will incorporate our work on assessing memento damage into WARCreate and WAIL.
- **Task 9** – WARC Viewer. To allow users to perform QA on created WARCs, we will develop a WARC viewer and a WARC images viewer to be integrated into WAIL.

Below, we detail our project activities and any changes to the original proposal related to each of these main tasks.

Task 1 - New WARCreate Modes

The record, countdown, and event modes were implemented into WARCreate and are available in a development version.

- *Record* mode - The logic for this feature already existed in WARCreate by design, even without enabling the feature, but the "history" of WARC data generated is cleared with each new page a user navigates to. The new interface controls whether this content should be cleared or not.
- *Countdown* and *event* modes – These additional modes only required that the extension reload the page or detect when the page has been reloaded.

In addition to implementing these new modes, we devoted significant time to refining the core operation of WARCreate and ensuring that pages are captured sufficiently.

One of the future ideas for WARCreate (outside of the scope of this grant) will be to transition it from a one-page-at-time capture mechanism to a truly browser-based crawler, capable of launching and capturing webpages based on a user-specified set of seeds.

Task 2 – WARC File Transfer Methods

As part of our Start-Up Grant, we added the ability for user-generated WARCs to be uploaded to a remote server. We have also successfully tested storing WARC files generated with WARCreate in a shared Google Drive folder and accessing the files through *wayback* instances on different machines. This allows for some level of sharing of the individual WARC files.

Even though the grant period has ended, our work on these file transfer methods will progress along with Mat Kelly's (PhD student funded by this grant) thesis work. Mat is developing a framework to allow access to both public and private/personal archives through a single interface. We believe that this will be a better method of allowing access for individually-generated WARCs than integrating these into existing public archives. Integrating with a public archive introduces the following issues:

- URL clashing - my facebook.com is different than your facebook.com, public archives may only have capture of login pages
- privacy - users will not want to upload private information potentially contained in the WARCs
- integrity - Archive-It should not be the arbiter of whether a WARC has been modified before upload

The Internet Archive and partners were awarded an IMLS grant (https://blog.archive.org/2015/10/08/imls-national-digital-platform-grant-awarded-to-advance-web-archiving/) to investigate APIs (WASAPI, shorthand for Web Archiving Systems API) for the transfer of WARCs and other archive files. The development of this API is ongoing and even after our NEH grant period is completed, we will work to integrate the WASAPI work into our tools as appropriate.

During this process, we have found that simple transfer of WARCs between systems may be more complicated than originally thought. Some tools that open WARC files tend to transform line breaks, which can break compliance with the WARC spec.

Task 3 – WAIL User Interface

Our collaborators at Michigan State reviewed all of the tools, including the WAIL user interface, and provided valuable feedback. As part of the evaluation process, our collaborators at Michigan State conducted a survey of the archiving needs of humanities scholars. The report is available as Appendix A.

Because of feedback and emerging technology that was presented during the course of the funding, we put significant efforts into overhauling and improving the operation of WAIL, not only the user interface.

The major updates we made to WAIL consisted of moving WAIL to Electron (http://electron.atom.io/), which allows us to develop a single codebase and deploy the system on Windows, Mac, and Linux platforms. We posted a write-up of our rationale in the conversion on our research blog (http://ws-dl.blogspot.com/2016/06/2016-06-03-lipstick-or-ham-next-steps.html).

In addition to moving to Electron, we shifted the *wayback* that was included in WAIL from OpenWayback to PyWb (https://github.com/ikreymer/pywb). PyWb provides for collection-based management of web archives, and the UI of WAIL was redesigned to be collection-centric akin to the Archive-It interface. Users can create collections and add seeds to a collection from the live web or from WARC/ARC files contained on their filesystem.

To support high-fidelity crawling of webpages, we developed an open-source library for browser-based archiving called node-warc (https://github.com/N0taN3rd/node-warc). We used this library to build Squidwarc (https://github.com/N0taN3rd/Squidwarc), a high-fidelity archival crawler that uses the Google Chrome browser (in headless mode) for rendering, so it can capture webpages that are JavaScript-heavy. Details of the development of node-warc and Squidwarc can be found in the "Replacing Heritrix with Chrome in WAIL, and the release of node-warc, node-cdxj, and Squidwarc" blog post at http://ws-dl.blogspot.com/2017/07/2017-07-24-replacing-heritrix-with.html

In the original version of WAIL, Heritrix was used as the primary archival crawler. In the Electron version of WAIL, we provide to the users the option to use Heritrix or WAIL-Archiver (a minimal version of Squidwarc) as the archival crawler by introducing several different archiving configurations:

- Page Only (WAIL-Archiver):
  - Preserve the page such that there is no difference when replaying the page from viewing the page in a web browser at preservation time
- Page + Same Domain Links (WAIL-Archiver):
  - Page Only option plus retrieve all links found on the page that come from the same domain as the page
- Page + All internal and external links (WAIL-Archiver):
  - Page + Same Domain Link option plus all links from other domains
- Heritrix Depth 1
  - Use the Heritrix crawler to archive the seed webpage and all webpages linked from the seed webpage
- Heritrix Depth 2
  - Use the Heritrix crawler to archive the seed webpage, all webpages linked from the seed webpage, and all webpages linked from those webpages (depth 2)

- Heritrix Depth 3
    - Heritrix Depth 2, plus all webpages linked from the Depth 2 webpages

We have also added Twitter-specific archiving features into WAIL. Users can sign into Twitter through WAIL and configure WAIL to monitor a Twitter users feed to be archived at specified intervals or to preserve tweets that match criteria which are also user-specified.

A video walk-through of WAIL's current features is available at https://www.youtube.com/watch?v=WTR9p9Vnab0

Task 4 – New Wayback Feature Integration into WAIL

We transitioned from OpenWayback to PyWb with WAIL-Electron. PyWb is a more recent implementation of *wayback* and is being currently maintained and updated. New features added into PyWb will easily be able to be integrated into WAIL.

We also included into WAIL MemGator (https://github.com/oduwsdl/memgator), a Memento aggregator, developed at ODU as part of a separate project. The inclusion of MemGator has allowed us to make progress on Task 7, providing local archive query from Mink.

We note that the Web Archive Player, developed by Ilya Kreymer, contains many of the features that we had planned for Task 9 - WARC Viewer. Because this has been released open-source, we can include it in WAIL without having to develop our own version.

Task 5 – Firefox Version of Mink

During the course of this project, it was announced (see https://blog.mozilla.org/addons/2015/08/21/the-future-of-developing-firefox-add-ons/) that Firefox will be adopting the same method for extensions as Chrome currently has. This resulted in Firefox extension APIs that we were familiar with from previous work being deprecated. We have continued to monitor the variations in the browser extension APIs and determined that our efforts were best put towards improving our core tools of WARCreate and WAIL and keeping the Chrome version Mink stable.

Task 6 – Mink User Interface

We performed significant improvements to the Mink user interface, based on feedback from our partners at Michigan State.

Versions of Mink prior to 2.0, which contains a user interface (UI) overhaul, relied solely on the UI injected by Mink into the web page's Document Object Model (DOM) tree. Mink's "status indicator" showed a rotating icon that stops rotating and reporting information once the process of querying the archives for captures was complete. The overhauled version of Mink uses Chrome "Browser Action" instead, in much the same way as WARCreate. In Mink version 2.0 and greater, the "processing" effect is now performed outside of the DOM in what is colloquially known as the toolbar. Once Mink finishes fetching the TimeMap (list of mementos) for the page currently being viewed on the live web, the memento count is displayed as a native Chrome badge on the Mink icon without any obtrusive UI shown to the user. Only when a user selects the Mink icon in the toolbar does Mink display its UI.

The Mink 2.x UI has been modernized compared to prior versions. Additionally, the UI can be retracted back to the toolbar by the user simply selecting the Mink icon again, giving a native toggle feel. Mink 1.x allowed the icon to be removed through a context (right-click) menu, which then was no longer present to re-show the Mink UI. The persistent browser action button in Mink 2.x retains this context in either state.

In addition to the new UI, Mink also adds a white-list and black-list feature that provides the user the ability to prevent Mink from processing URIs or URI schemes (wild cards and/or regular expression URI matching are allowed) where the memento count does not make as much sense, e.g., a user's Gmail inbox.

Task 7 – Local Archive Query in Mink

Local archive query is currently part of Mat Kelly's dissertation work. He is developing a framework for how Memento-based tools can aggregate resources from both public and private web archives. Work on this task continues with Mat's research and will continue even after the grant period has ended, as this is a major part of Mat's dissertation.

Task 8 – Memento Damage Assessment

Our work on memento damage assessment is continuing on another project. We currently have a web service available for testing at http://memento-damage.cs.odu.edu/. Once this work is complete, we will incorporate it into WAIL to allow users to assess how well their pages were archived.

Task 9 – WARC Viewer

As mentioned in the update for Task 4, Web Archive Player, developed by Ilya Kreymer, contains many of the features that we had planned for Task 9 - WARC Viewer. Ilya is currently working on a re-implementation of the tool based on his https://webrecorder.io/ toolset. We will not be developing our own tool to prevent duplication of effort. If needed, we plan to incorporate his new tool into WAIL.

Other Tasks

In order to help facilitate our collaboration with our Michigan State partners, we wrote a summary of the capabilities of the three tools and their use cases. The tool summary is available in Appendix B, and the use cases are in Appendix C.

Accomplishments

We are proud of the advancements that we helped to make in the state-of-the-art in high-fidelity web archiving. Our work on WAIL, and especially the Electron version of WAIL, helped to highlight the need for browser-based archiving. Other browser-based archiving tools such as webrecorder.io and Browsertrix (https://github.com/ikreymer/browsertrix) have been developed since our project began. WAIL and WARCreate are still the only ones that emphasize personal and private archiving of webpages. Our team members have also contributed to webrecorder.io and collaborated with the developers of that tool.

The web archiving landscape has changed significantly since we started this project in 2014. Some of the project goals and tasks were made unnecessary by new technologies and the rise of different modes of web archiving. The early prototypes of our tools helped to inspire new tools built by professional developers. For instance, webrecorder.io began in May 2014, after we had released our early versions of WARCreate and WAIL (initial developments were supported by an NEH DH start-up grant).

Audiences

The intended audience for our tools is a wide range of scholars and individuals. A common feature is that these tools enable individuals, not institutions, to perform web archiving, so ease of use is of importance.

Evaluation

At the beginning of the project, our partners at Michigan State conducted a survey regarding archiving practices. The result of this survey is included in Appendix A.

In retrospect, we probably proposed too many tasks. Especially with the changing browser and archiving landscape over the past three years, we spent considerable effort in ensuring that the tools were compatible with the latest operating systems and versions of browsers. We also followed the developments of other projects to ensure that we were not duplicating efforts. Even with the development of tools like webrecorder.io, our tools still offer users a method to privately capture webpages. Using WARCreate or WAIL, captures never leave the user's local computer.

The tools produced in this project were not formally evaluated, but several public releases were made and feedback was solicited from our partners, advisory board, and the general public. We presented the tools developed in this project in various forums (see "Grant Products" for a full list). Through these forums, email, and GitHub comments, we received positive feedback from the public.

Users especially appreciated that the tools required little setup and configuration, were cross-platform, easy-to-use, and had an intuitive user interface. In terms of operation, users liked that WAIL could be used to view WARCs that were generated by other tools and could be used to archive Twitter. Several users saw the privacy of WAIL and WARCreate as a strong advantage. No data is sent to an outside server or cloud service; all WARCs are retained locally.

Users had several feature requests for WAIL, including the ability to enter multiple URIs at a time for a crawl, the ability to configure Heritrix to work with a proxy, and improved event tracking and logging. For WARCreate, some users requested that the compressed WARC.gz format be supported and that raw HTTP responses be retained (these are not currently accessible from the Chrome Extensions API).

Continuation of the Project

Our current plans to continue the project are in support of our students' academic progress.

Mink will likely be modified by PhD student Mat Kelly as a demonstration of his dissertation work to aggregate public and private archives. Kelly can also use both WAIL and WARCreate to create private archives for use in his demonstrations.

MS student John Berlin will be continuing in the PhD program at ODU and will continue to investigate issues related to high-fidelity web archiving. Some of his dissertation work may include improvements to WARCreate and WAIL.

Several of our other research projects involve creating private web archives and students have been using WARCreate and WAIL for this purpose.

Long Term Impact

Through this project, we believe that we have amplified the need for high-fidelity, browser-based archiving and have highlighted the need for private and personal web archives.

Grant Products

Our grant products consist of software, publications, blog posts, and presentations.

Software

All of our tools are available at https://oduwsdl.github.io/nehdhig2017/. The links to the individual github repositories are listed below:

- WAIL (Web Archiving Integration Layer), http://machawk1.github.io/wail/
- WARCreate Google Chrome extension, https://github.com/machawk1/warcreate
- Mink Google Chrome extension, https://github.com/machawk1/mink
- WAIL-Electron, https://github.com/n0tan3rd/wail
- Squidwarc, https://github.com/N0taN3rd/Squidwarc

- Node-warc, https://github.com/N0taN3rd/node-warc
- Node-cdxj, https://github.com/N0taN3rd/node-cdxj

Publications

- Mat Kelly, Michael L. Nelson, and Michele C. Weigle, "Mink: Integrating the Live and Archived Web Viewing Experience Using Web Browsers and Memento", Proceedings of the IEEE/ACM Joint Conference on Digital Libraries (JCDL), June 2014, doi: 10.1109/JCDL.2014.6970229 http://www.cs.odu.edu/~mln/pubs/jcdl-2014/jcdl-2014-kelly-mink.pdf, http://www.slideshare.net/matkelly01/mink-integrating-the-live-and-archived-web-viewing-experience-using-web-browsers-and-memento
- John Berlin, Mat Kelly, Michael L. Nelson, and Michele C. Weigle, "WAIL: Collection-Based Personal Web Archiving", Proceedings of the IEEE/ACM Joint Conference on Digital Libraries (JCDL), June 2017, doi: 10.1109/JCDL.2017.7991619 http://www.cs.odu.edu/~mkelly/papers/2017_jcdl_wail.pdf

Blog Posts

- NEH ODH Project Directors' Meeting Trip Report, September 15, 2014, http://ws-dl.blogspot.com/2014/09/2014-09-17-neh-odh-project-directors.html
- Archive-It Partners Meeting Trip Report, November 18, 2014, http://ws-dl.blogspot.com/2014/11/2014-11-20-archive-it-partners-meeting.html
- "Archive What I See Now" Project Funded by NEH Office of Digital Humanities, July 22, 2014, http://ws-dl.blogspot.com/2014/07/2014-07-22-archive-what-i-see-now.html
- Lipstick or Ham: Next Steps for WAIL, June 3, 2016, http://ws-dl.blogspot.com/2016/06/2016-06-03-lipstick-or-ham-next-steps.html
- Electric WAILs and Ham, February 13, 2017, http://ws-dl.blogspot.com/2017/02/2017-02-13-electric-wails-and-ham.html
- Replacing Heritrix with Chrome in WAIL, and the release of node-warc, node-cdxj, and Squidwarc, July 24, 2017, http://ws-dl.blogspot.com/2017/07/2017-07-24-replacing-heritrix-with.html

Presentations

- "Archive What I See Now", September 15, 2014, lightning talk at NEH Project Directors' Meeting, http://www.slideshare.net/mweigle/2014-weigleneh
- "Tools for Managing the Past Web", November 18, 2014, presented at Archive-It Partners Meeting, http://www.slideshare.net/mweigle/2014-weigleaitpublic
- "Tools For Managing the Past Web", February 20, 2015, presentation at ODU ECE Seminar, http://www.slideshare.net/mweigle/2015-oduecetalk
- "Web Archiving Activities of ODU's Web Science and Digital Library Research Group", talk at The National Symposium on Web Archiving Interoperability, February 21, 2017, http://ws-dl.blogspot.com/2017/03/2017-03-02-national-symposium-on-web.html
- "Web Archiving Integration Layer (WAIL)", demo at the Personal Digital Archiving conference, March 29-31, 2017, demo video at https://www.youtube.com/watch?v=hOfOsHW2lU4
- "Archive What I See Now - Personal Web Archiving with WARCs", presentation at the 2017 International Internet Preservation Consortium (IIPC) Web Archiving Conference, June 14-16, 2017, https://www.slideshare.net/machawk1/archive-what-i-see-now-personal-web-archiving-with-warcs

Appendix A – Archiving Survey Report

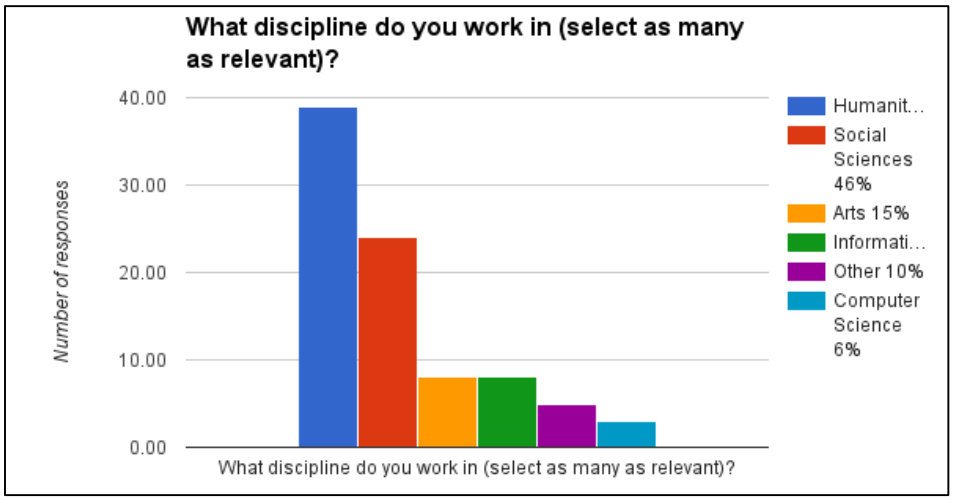
Abstract
After surveying 58 participants across multiple disciplines about their archiving practices and preferences, we learned that a majority of the participants value tools that allow them to archive content as needed, in a way that allows them the utmost privacy.

Survey questions
In order to offer recommendations for fine tuning the archiving tool suite, we constructed a 15 question survey and sent it out to researchers via discipline-specific listservs. We hoped to learn researchers current archiving methods, their preferences for archiving tools, and their preferences for storing archived content content.

1. Do you archive internet-based content?
2. What tools do you use to archive internet-based content?
3. Do you routinely archive content in batches (over a few hours) or as needed?
4. If you archive content, do you typically do so for individual research purposes or collaborative research purposes?
5. What kinds of content do you archive?
    a. Popular social media?
    b. Games?
    c. News and/or Information sites?
    d. Video, Photo, and/or Music sites?
    e. Private communication?
6. When you archive content, is it primarily publicly available or privately available?
7. A team of researchers is developing a new tool to archive internet content. Would you want this tool to constantly sit in your browser, giving you the ability to view archived pages?
8. Would you prefer this tool to be in the extensions/plugins/address bar section of your browser, rather than on the webpage itself?
9. Would you want the ability to turn this tool on and off?
10. Would you want your content publicly stored, or would you want to store your data privately?
11. What discipline do you work in?

Demographics of Survey Participants
After reviewing the data that we collected, we learned that participants who took our survey were mostly from the humanities and social sciences and their archiving practices occur on an as needed basis, mostly focusing on news sites, multimedia sites, and social media.

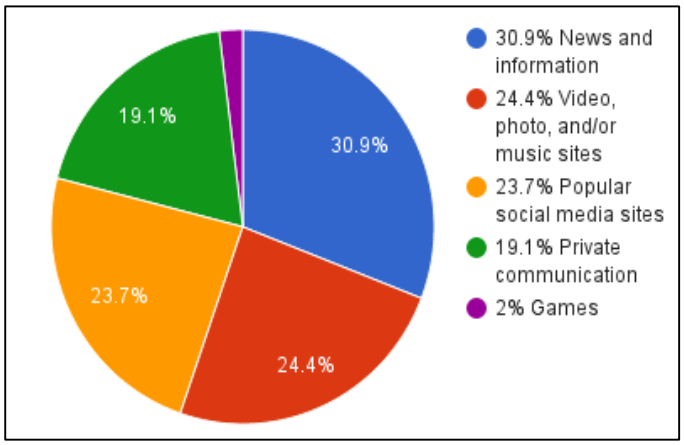
Discussion
Participant's disciplines
Overall we received 58 responses to our survey. Most of the participants identified their discipline as humanities or social sciences, while only some of the participants identified their disciplines as Art, Information Sciences, and Computer Sciences.



Participant's current archiving practices

When asked whether they archived content in batches (over a few hours) or as needed, a majority of participants (88.2%) responded that they archived content as needed, while only 11.8% answered that they archived in batches. We imagine that knowing this archiving practice will help fine-tune the development of the archiving suite by rethinking some of the use cases.

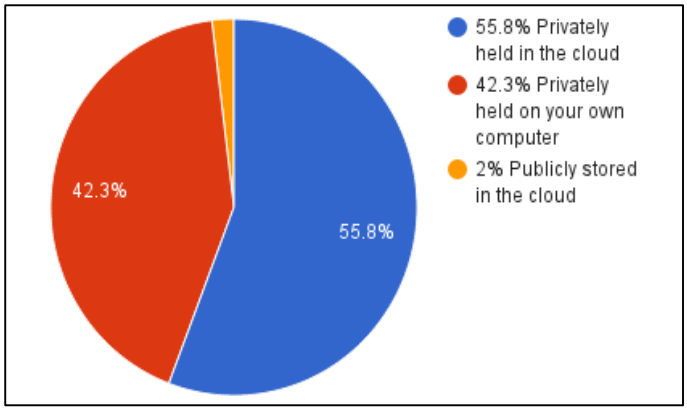
When asked what kinds of content do you archive, participants were extremely diverse, answering that they archived news and/or information sites (30.9%), popular social media sites (23.7%), video, photo, and music sites (24.4%), and private communication (19.1%). Knowing that participants are looking to archive social media in these numbers makes configuring the tools for these sites a priority as development continues. For example, 12% of researchers who archive social media said that they archived content from Pinterest.



Storage and privacy preferences

When asked where they would want their content stored, participants overwhelmingly chose private options (on cloud 55.8%, on computer 42.3%).

When asked about preferred features for a new archiving tool, 63% of participants said they would want a tool that sits in their browser and 84% said they would rather the tool be in the extension/plugins/address bar. A majority of participants (92.2%) want the ability to turn the tool off.



Appendix B – Tool Summary

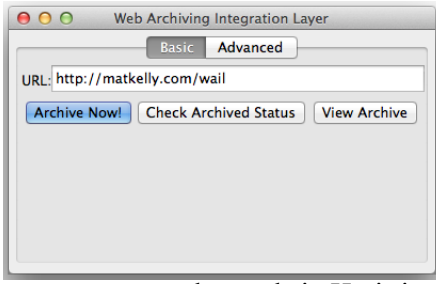
WARCreate

WARCreate is a browser extension that can create an archive of a single web page in the standard WARC format and save it to local disk.

WARCreate can capture any web page from the perspective of the web browser. This means that web pages requiring authentication, pages from social media sites, pages displayed after some user interaction, and page elements such as ads generated based on the context of the particular web browser can all be archived in a standard format. This type of archiving is not available with standard web archiving tools, such as Heritrix (the crawler used by the Internet Archive).

WAIL

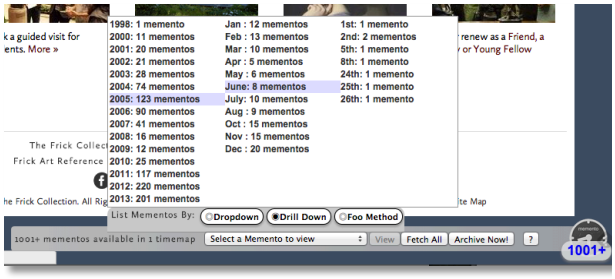
WAIL (Web Archiving Integration Layer) is a stand-alone desktop application (Windows and MacOS X) that provides one-click installation and GUI-based configuration of both Heritrix and *wayback* (archive replay tool used by the Internet Archive) on the user's personal computer. This allows users to view the WARCs that they created with WARCreate in an interface similar to the Internet Archive's Wayback Machine, with which they may already be familiar.



The "Archive Now!" button uses Heritrix to set up a crawl of the specified webpage and puts the created WARC in the proper location to be indexed by *wayback*. The "Check Archive Status" button will show the status of the current crawl (crawling, indexing, available in *wayback*). Once the WARC has been indexed, the "View Archive" button will bring up the local *wayback* interface for viewing all archives for the URL.

In addition, WAIL includes advanced options that would allow the user to set up web crawls in Heritrix to be repeated at a certain intervals and at certain depths (levels of links to follow). The advanced options panel also provides service management/status, access to the archives folder directory (i.e., where the WARCs are stored), direct access to modify the configuration files produced for *wayback* and Heritrix, etc.

Mink



Mink is a browser extension for Google Chrome that provides access to archived versions of live webpages. In the future, Mink will be able to access locally stored WARC files created by WARCreate.

When using the Mink extension, if a live webpage has archived versions accessible by the Memento protocol, an icon will appear in the lower right corner of the webpage. When the user clicks on the icon, the Mink tool bar will expand, showing the number of mementos available for that URI. The user can then select the desired datetime from either a drop-down list or date-based navigation columns. The selected memento will be loaded in the browser. The user can continue to browse the page in the past or click the "Return to Live Web" button.

Mink also allows users to submit the loaded webpage to one of several public archiving services, such as the Internet Archive and Archive.today, through the "Archive Now!" button.

Appendix C – Use Cases

General

With these three tools, a researcher could, in her normal workflow, discover a web resource (using her browser), locally archive the resource as she saw it (using **WARCreate** in her browser), and then later index and replay the archived resource (using **WAIL**). Once the archived resource is indexed, it would be available for view in the researcher's browser using **Mink** (*in future work*). If the web page is public, **Mink** allows the user to submit the page for archiving by one of several web archiving services.

WARCreate and WAIL

A researcher wants to locally archive a webpage for sharing with others. She loads the webpage in her browser (with the WARCreate extension enabled) and clicks the WARCreate icon. The "Generate WARC" button is displayed. She clicks the button and a WARC is generated and saved locally on her computer. Currently, the WARC is saved to the browser's default Downloads location.

To share the file, she can either email it to a colleague or share it via any file sharing service such as Dropbox or Google Drive.

To view the WARC file, the researcher, or her colleague, opens the WAIL desktop application and clicks the "View Archived Files" button to open the folder of WARCs. The user drags the newly created WARC into this folder for indexing. In the Basic interface, the user clicks the "View Archive" button to open WAIL's wayback interface in their web browser. Then, the user can enter the URL and view the indexed mementos, as with the standard Wayback Machine.

WAIL can also be used to set up advanced crawls using Heritrix.

Mink

Currently, Mink is more about browsing the archives while you're browsing the live web and is not yet directly tied to the other two tools.

A researcher is viewing cnn.com or some other live webpage. She wants to access a version of the page from the previous day. The Mink icon indicates that there are archived versions, or mementos, of the page available (through the number displayed on the Mink icon). The researcher clicks the Mink icon and is presented with a toolbar allowing for navigating through the available mementos for the page. When she chooses a datetime, that memento will be displayed in her browser. She may continue to browse the past web using links in the memento, or she can return to browsing the live web by clicking the Mink icon and pressing the "Return to Live Web" button.

If the live webpage that the researcher is interested in has no recent mementos and she feels that the page should be archived, she can use Mink to submit the webpage to a public web archiving service. She clicks the Mink icon and then presses the "Archive Now!" button to reveal a set of archives that the page can be submitted to. She chooses the Internet Archive (Archive.org) and the URL of the currently displayed page is submitted for archiving. Note that any interaction with the web archiving service is independent of the Mink extension (for instance, Archive.today displays log information as it is archiving the requested page). Also note that the web archiving services may not be able to archive the exact state of the webpage as seen in the browser (as WARCreate can).